

Introducción a la Filoinformática: Pan-genómica y Filogenómica microbiana – <https://github.com/vinuesa/TIB-filoinfo>

Pablo Vinuesa ([@pvnmex](mailto:vinuesa@ccg.unam.mx))  
Centro de Ciencias Genómicas, (CCG-UNAM), Campus Morelos, Cuernavaca, México <http://www.ccg.unam.mx/~vinuesa/>

Todo el material del curso (presentaciones, tutoriales y datos) lo encontrarás en: <https://github.com/vinuesa/TIB-filoinfo>

• Tema 1: conceptos básicos de evolución molecular y genómica microbiana

- Relación entre filogenética y evolución molecular
- Concepto y tipos de homología
- Árboles de genes vs. árboles de especies
- Fundamentos de genómica microbiana y consideraciones para la selección de marcadores
- Marcadores moleculares y tasas de evolución
- Protocolo básico de análisis filogenético
- Clasificación de métodos de inferencia filogenética
- Tipos de árboles y su enraizamiento
- Reconstrucción y evolución de caracteres sobre filogenias

## Parte I: Introducción a la Inferencia Filogenética

### Conceptos básicos:

- \* filogenia, evolución molecular, homología y (pan)genómica microbiana
- \* tasas de evolución y selección de marcadores moleculares

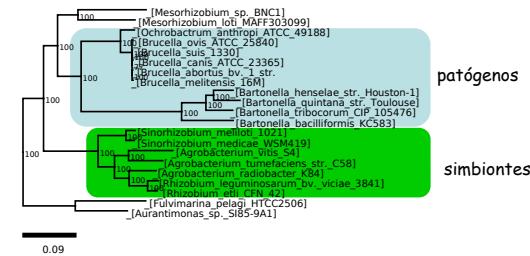
### Libros de referencia recomendados:

- Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, INC., Sunderland, MA.
- Futuyma, D.J. 2017. Evolution, 4th Ed. Sinauer Associates, INC., Sunderland, MA.
- Graur, D., Li, W.H. 2000. Fundamentals of Molecular Evolution. Sinauer Associates, Inc., Sunderland.
- Lemey P., Salemi M., Vandamme A.-M. 2010. The phylogenetic Handbook. Cambridge Univ. Press. UK
- Keith, Jonathan M. (Ed.) Bioinformatics: Volume I: Data, Sequence Analysis, and Evolution (Methods in Molecular Biology) <https://www.springer.com/gp/book/9781493966202>
- Nei, M., Kumar, S., 2000. Molecular Evolution and Phylogenetics. Oxford University Press, Inc., NY.
- Page, R.D.M., Holmes, E.C., 1998. Molecular Evolution - A Phylogenetic Approach. Blackwell Science Ltd, Oxford.
- Yang, Z., 2014. Molecular Evolution – a statistical approach. Oxford University Press, Oxford, UK

## La relación entre filogenética y evolución molecular:

- La filogenética tiene por objetivo el trazar la relación ancestro descendiente de los organismos (árbol filogenético) a diferentes niveles taxonómicos, incluyendo el árbol universal, haciendo una reconstrucción de esta historia de ancestría en base a **caracteres homólogos**, tanto **morfológicos** como **moleculares**.

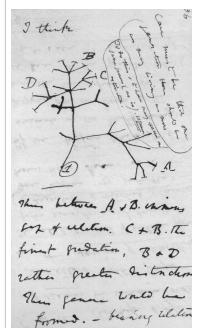
Las hipótesis filogenéticas resultantes son la base para hacer **predicciones** (inferencias) sobre propiedades biológicas de los grupos revelados por la filogenia mediante el mapeo de caracteres sobre la topología (hipótesis evolutiva). También proveen el contexto comparativo para poder inferir patrones de **evolución molecular**.



## Evolución de la filogenética como disciplina científica



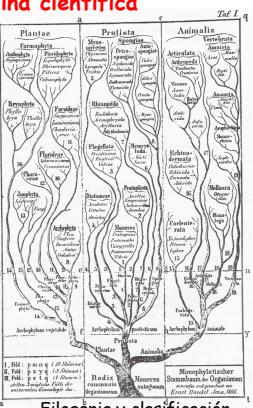
Los primeros intentos de reconstruir la historia filogenética no estaban basados en criterios objetivos.



Reflejaban las ideas o hipótesis plausibles generadas por expertos de grupos taxonómicos particulares.

La mayor parte de la 1a. mitad del SXIX los sistemáticos estaban más preocupados por el problema de definir a las especies biológicas, descubrir mecanismos de especiación y la variación geográfica de las especies, que en entender su filogenia.

No fue hasta los 40's y 50's del siglo pasado que los esfuerzos de **Walter Zimmermann** y **Willi Henning** comenzaron a definir métodos objetivos para reconstruir filogenias en base a caracteres compartidos entre organismos fósiles y contemporáneos.



Filogenia y clasificación de la vida tal y como la propuso Ernst von Haeckel en 1866



## ¿Porqué estudiar filogenética y evolución molecular?

Corolario I:

"Nothing in biology makes sense except in the light of evolution"

- Theodosius Dobzhanski, 1973

(*The American Biology Teacher* 35:125)

Corolario II:

"Nothing in evolutionary biology makes sense except in the light of a phylogeny"

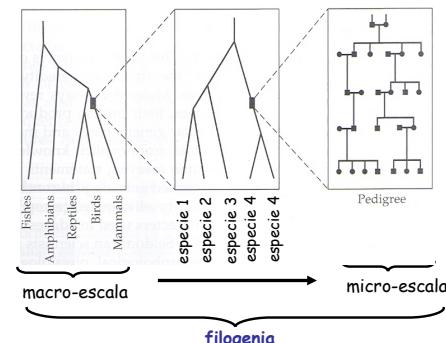
- Jeff Palmer, Douglas Soltis, Mark Chase, 2004

(*American J. Botany* 91: 1437-1445)



## El concepto de filogenia y homología: definiciones básicas

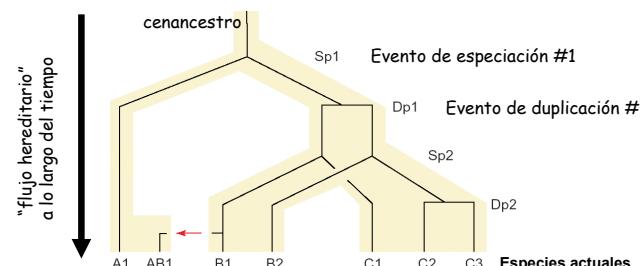
"The stream of heredity makes phylogeny; in a sense, it is phylogeny. Complete genetic analysis would provide the most priceless data for the mapping of this stream".  
G.G. Simpson (1945)



**Filogenia:** historia evolutiva del flujo hereditario a distintos niveles evolutivos/temporales, desde la genealogía de genes en poblaciones (micro-escala; dominio de la genética de poblaciones) hasta el árbol universal (macro-escala)

## El concepto de homología: definiciones básicas

Subtipos de homología: ortología, paralogía y xenología



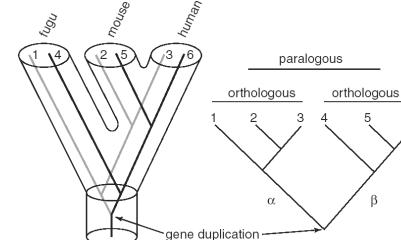
**ortología:** relación entre secuencias en la que la divergencia acontece tras un evento de especiación. El ancestro común es el cenáncestro. La filogenia recuperada de estas secuencias refleja la filogenia de las especies.

**paralogía:** condición evolutiva en la que la divergencia observada acontece tras un evento de duplicación génica. La mezcla de ortólogos y parálogos en un mismo análisis filogenético recupera la filogenia correcta de los genes pero no necesariamente la de los organismos o taxa.

**xenología:** relación entre secuencias dada por un evento de transferencia horizontal entre linajes. Distorsiona fuertemente la filogenia de las especies.

## Arboles de genes vs. árboles de especies - el problema de la definición de relaciones de homología

- **La filogenia de especies** puede ser inferida erróneamente cuando se reconstruye en base a secuencias parálogas y no se muestran todas las copias (p. ej. si muestreamos sólo las copias 1, 3 y 5 eq ((fugu, human), mouse) !!!
- Más compleja aún es la estima de la filogenia de especies si ha habido pérdida diferencial de parálogos en los distintos linajes a comparar
- Por tanto la inferencia de una filogenia de especies se realizará preferentemente usando ortólogos de copia única



### Filogenias de genes vs. filogenias de especies

**El coalescente multiespecies**

- Cada gen tiene su propia tasa de evolución
- Los genes de un organismo pueden tener historias discordantes entre sí
- Por tanto, las filogenias de genes no reflejan necesariamente la filogenia de las especies

### La TGH y la filogenia universal – problemas y limitaciones evidenciadas desde la perspectiva genómica

**Filogenia genómica – ¿árbol o red?**

**Problemas e incógnitas de las filogenias profundas:**

- ¿Cuántos ortólogos son realmente universales?
- Distinción entre ortólogos, parálogos y xenólogos
- ¿Cuánta señal filogenética queda en las secuencias?
- ¿Alineamientos confiables?
- Métodos de reconstrucción y artefactos
- Congruencia de señales filogenéticas provenientes de distintos genes
- ¿Existió realmente un sólo ancestro?
- ... una larga lista

### TGH y el pan-genoma microbiano - una perspectiva de genómica comparada

**Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli***

R. A. Welch\*, V. Burland†‡, G. Plunkett III†, P. Redford\*, P. Roesch\*, D. Rasko§, E. L. Buckles§, S.-R. Liou†, A. Boutin†\*\*, J. Hackett†, G. Stroud†, G. F. Mayhew†, D. J. Rose†, S. Zhou†, D. C. Schwartz†, N. T. Perna§§, H. L. T. Mobley§, M. S. Donnenberg§, and F. R. Blattner†

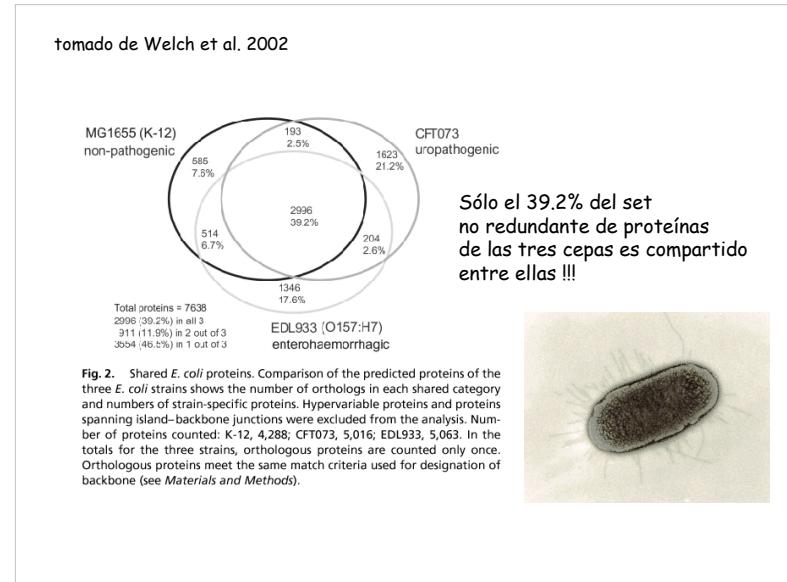
\*Department of Medical Microbiology and Immunology, Laboratory of Genetics, †Genome Center of Wisconsin, and §Animal Health and Biological Sciences, University of Wisconsin, Madison, WI 53706, and ‡Department of Microbiology and Immunology, †Division of Infectious Diseases, Department of Medicine, University of Maryland School of Medicine, Baltimore, MD 21201

Edited by John J. Mekalanos, Harvard Medical School, Boston, MA, and approved October 24, 2002 (received for review August 30, 2002)

17020–17024 | PNAS | December 24, 2002 | vol. 99 | no. 26

Se comparan las estructuras genómicas de tres cepas de *Escherichia coli*

*E. coli* uropatogénica CFT073  
*E. coli* enterohemorrágica EDL933  
*E. coli* comensal K12



tomado de Welch et al. 2002

### Islands and evidences for their horizontal transfer

- 85% (52 out of 61) of island codons have a significantly different codon usage vs. backbone
- EDL and CFTshare only 10% of island genes, but >98% identity among backbone encoded proteins in the 3-strain comparison
- CFT and EDL strains contain 60 and 57 islands >4 kb, most of them at the same relative positions with respect to backbone markers, although island contents are unrelated
- 12 CFT073 and 10 EDL933 islands are closely associated to tRNA genes

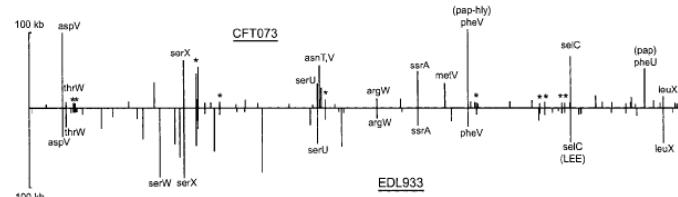


Fig. 3. Locations and sizes of CFT073 and EDL933 islands. Island size, vertical axis; position in colinear backbone, horizontal axis. All islands >4 kb are shown. Islands located at tRNAs are indicated by tRNA labels. One tmRNA (ssrA) is also an insertion target. \*, CFT073 and EDL933 islands in the same backbone location but not near tRNAs.

### Corolario:

A la luz de estos resultados queda evidenciado que ha de ejercerse un gran cuidado a la hora de seleccionar genes para la reconstrucción de filogenias de especies bacterias.

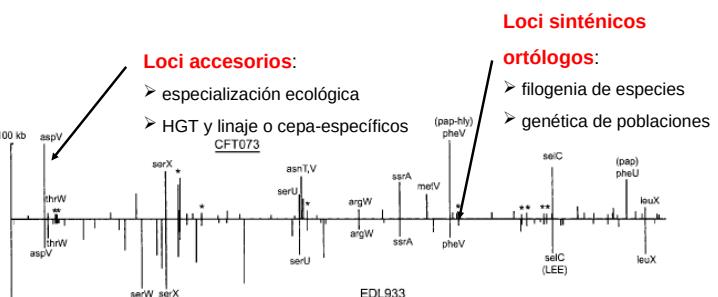


Fig. 3. Locations and sizes of CFT073 and EDL933 islands. Island size, vertical axis; position in colinear backbone, horizontal axis. All islands >4 kb are shown. Islands located at tRNAs are indicated by tRNA labels. One tmRNA (ssrA) is also an insertion target. \*, CFT073 and EDL933 islands in the same backbone location but not near tRNAs.

### El pangenoma microbiano: pangenomas abiertos y cerrados genoma core y accesorio

The microbial pan-genome  
 Duccio Medini<sup>1</sup>, Claudio Donati<sup>1</sup>, Hervé Tettelin<sup>2</sup>, Vega Masignani<sup>1</sup> and Rino Rappuoli<sup>1</sup>

Current Opinion in Genetics & Development 2005, 15:589–594

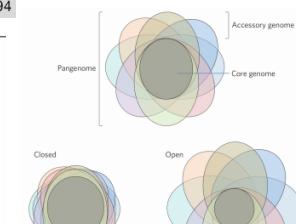
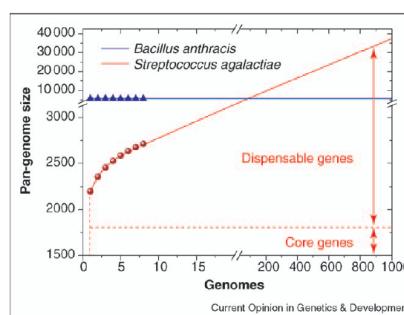


Figure 1 | Schematic representation of pangenomes as Venn diagrams. Species differ in the sizes of their pangenomes, with larger, more open pangenomes correlating with larger long-term effective population sizes and the ability to migrate.

McInerney et al. 2017 Nat. Micro.

### Inferencia filogenética: los detalles (básicos)

Datos moleculares en filogenética: tasas y restricciones evolutivas

El protocolo básico para un análisis filogenético de secuencias moleculares

Determinación de homología entre múltiples secuencias: slineamientos múltiples

Clasificación de métodos filogenéticos

Tipos de árboles filogenéticos y su "anatomía"

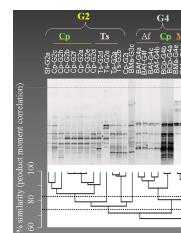
Evolución de caracteres

## Marcadores moleculares usados en filogenética y evolución molecular

### Polimorfismos de DNA y proteínas

#### I) Marcadores dominantes (# secuencias)

- RFLPs
- Fingerprints genómicos (AFLPs, RAPDs, Rep-PCR, SINEs SSCP, NSNPs ...)
- Análisis multilocus de isoenzimas
- etc ...



Los datos moleculares revelan información genética. Sólo datos con una base genética son de interés en filogenética y evolución. De ahí que los marcadores moleculares son generalmente los favorecidos para hacer inferencias filogenéticas y evolutivas a distintos niveles taxonómicos.

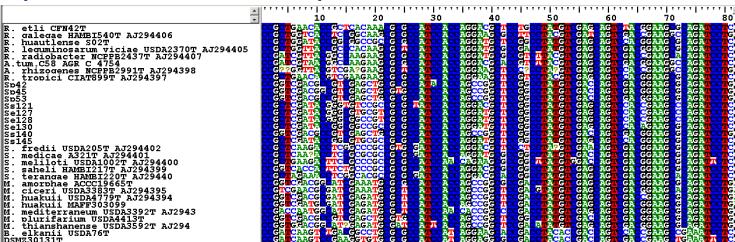
Los caracteres fenotípicos muchas veces tienen una base genética menos clara y están gobernados por las interacciones de muchos genes con el ambiente. Muchos fenotipos presentan gran plasticidad, es decir, que un mismo genotipo puede presentar una gradación de fenotipos. Esta variación fenotípica puede confundir las verdaderas relaciones filogenéticas y determinación de parentescos.

El uso de protocolos de PCR permite acceder a todo el mundo biológico para scrutinios genéticos

Los métodos moleculares permiten una fácil y robusta distinción entre homología y analogía y permiten hacer comparaciones de divergencia evolutiva usando métricos universales

## Marcadores moleculares usados en filogenética y evolución molecular

### II) Secuencias moleculares DNA/proteína



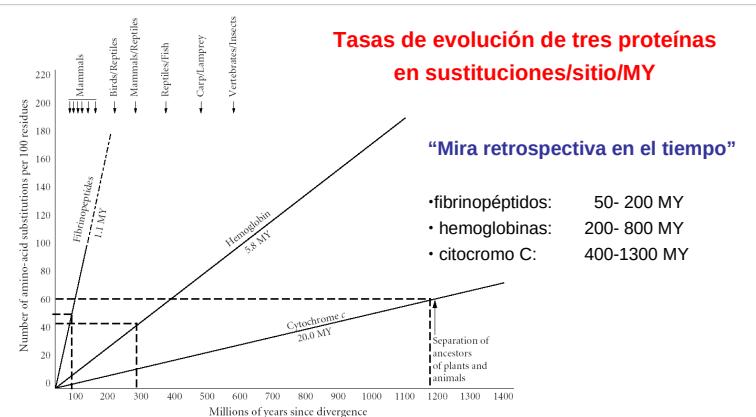
- La premisa fundamental en evol. molec. es que en dichas secuencias se encuentra escrita una buena parte de su historia evolutiva.
- Secuencias de DNA representan el “nivel anatómico” más fino de un organismo
- Buena parte de la biología moderna tiene por objetivo revelar la información contenida en secuencias moleculares
- Para inferir la historia de relaciones de ancestría entre un conjunto de secuencias homólogas hemos de **determinar las correspondencias de homología entre los caracteres** haciendo un **alineamiento múltiple de las secuencias**

## Selección de marcadores adecuados para hacer inferencias evolutivas a distintos niveles de profundidad filogenética

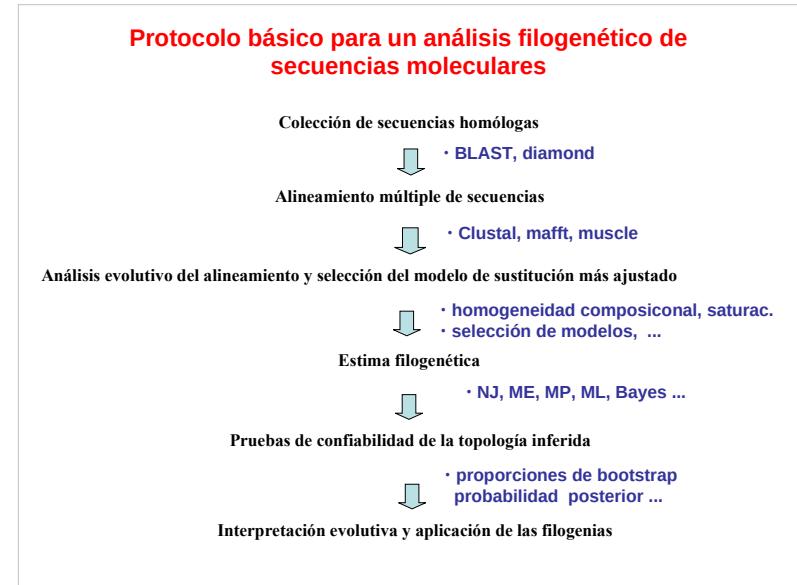
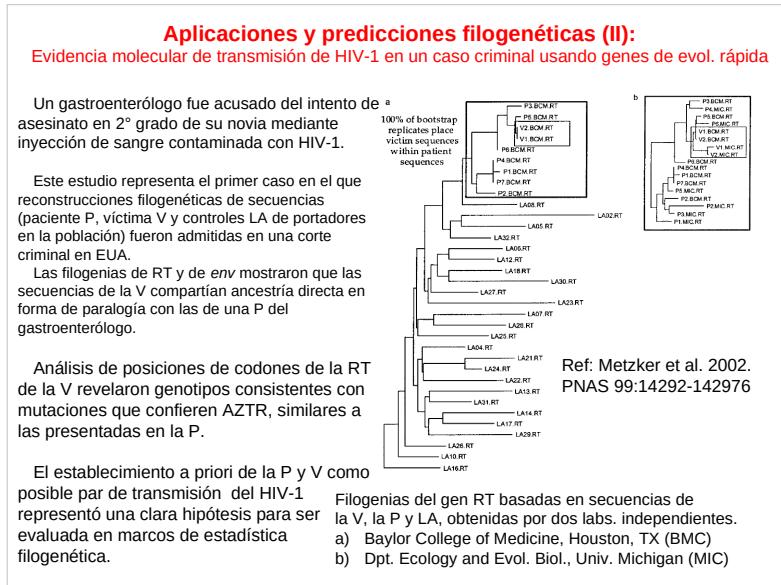
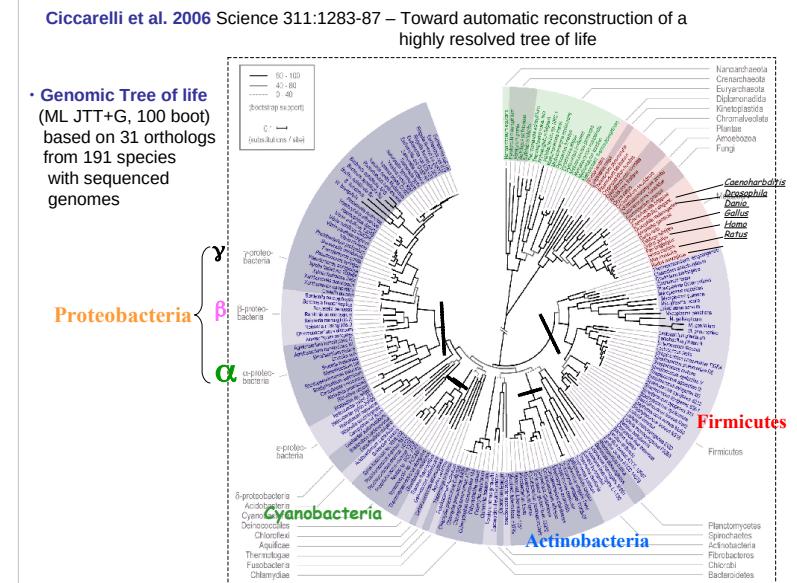
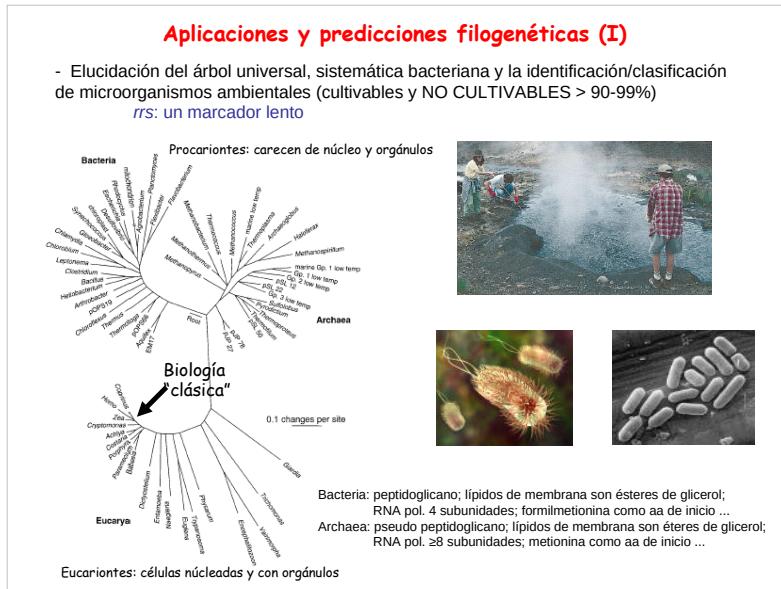
### Restricciones funcionales vs. tasas de sustitución:

- Existe gran variabilidad en la tasa de sustitución entre genes y dominios génicos:
  - intrones vs. exones
  - regiones codificadoras vs. regiones intergénicas o pseudogenes
  - residuos catalíticos vs. no catalíticos, dominios estructurales vs. no estructurales
  - 3as. posiciones vs. 1as y 2as en codones de secuencias codificadoras,
  - asas vs. orquillas en rRNAs y tRNAs ...
- Existen genes de evolución muy rápida o muy lenta:
  - fibrinopéptidos evolucionan una tasa x900 > a la de ubiquitina y x20 > citocromo C
  - genes de HIV evolucionan a  $\times 10^9$  veces la tasa de un gen humano promedio!
- Tasas de evolución y la teoría neutral de evolución molecular:

el reloj molecular, calibración y datación de eventos de especiación/extinción de linajes y de pandemias ...

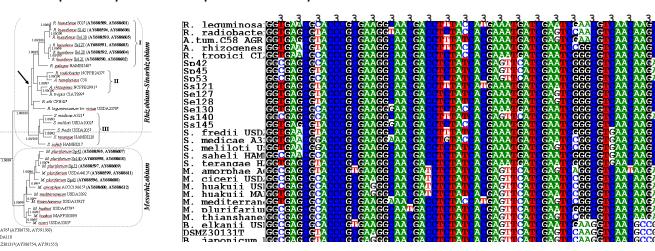


- Distintas proteínas presentan diversas tasas de sustitución. Así los fibrinopéptidos presentan relativamente pocas restricciones, presentando una elevada tasa de sustitución neutral. Citocromo C, en cambio, presenta mayores restricciones evolutivas y presenta una tasa de sustitución menor. La hipótesis del reloj molecular dice que esta tasa, para ciertas proteínas, es constante en distintos linajes.
- (de Hartl y Clark, 1997. Principles of Population Genetics, Sinauer)



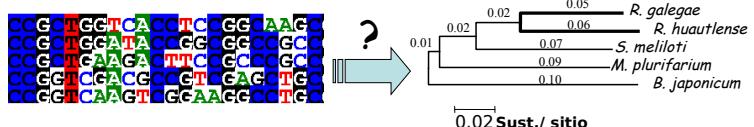
## Homología entre secuencias de DNA: alineamientos múltiples

- A lo largo de la evolución las secuencias descendientes de otra ancestral van acumulando diversos tipos de mutaciones. Estas son **mutaciones puntuales** o **reorganizaciones genómicas**, que pueden involucrar  **inserciones, delecciones, inversiones, translocaciones o duplicaciones**, mediados por distintos mecanismos de recombinación (homóloga e ilegítima)
- Cualquier análisis filogenético y/o evolutivo de secuencias moleculares requiere de un **alineamiento** para poder comparar sitios homólogos entre las secuencias a estudiar. Para ello se escriben las secuencias en filas una sobre la otra, de modo que los sitios homólogos quedan alineados por columnas. Cada sitio o columna del alineamiento corresponde a un **caracter**, y los nt o aa que ocupan dichas posiciones representan los distintos **estados del carácter**



## Inferencia Filogenética - introducción

- La **inferencia de relaciones filogenéticas** a partir de secs. moleculares requiere de la selección de uno de los muchos métodos disponibles
- Con frecuencia la inferencia filogenética es considerada como una "caja negra" en la que "entran las secuencias y salen los árboles" (filogenias estimadas con MEGA)



### Objetivos de esta taller son:

- desarrollar un marco conceptual para entender los fundamentos teóricos (filosóficos) que distinguen a los distintos métodos de inferencia (clasificación de métodos)
- presentar el uso de **modelos y suposiciones** en filogenética
- manejo empírico de diversos paquetes de software para inferencia filogenética bajo diversos criterios

## Inferencia filogenética molecular - clasificación de métodos

- Podemos clasificar a los métodos de reconstrucción filogenética en base al **tipo de datos** que emplean (**caracteres discretos vs. distancias**) y si usan un **método algorítmico** o **un criterio de optimización** para encontrar la topología

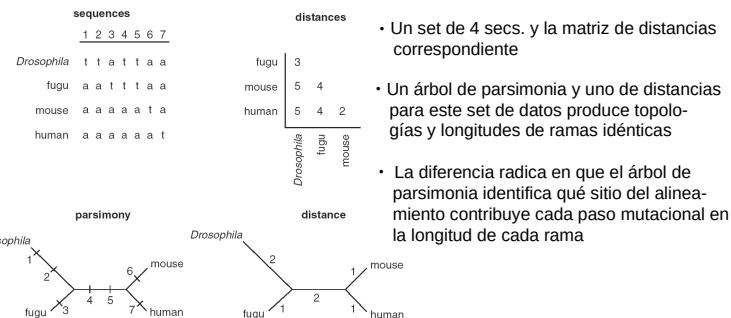
### Tipo de datos

Método de reconstrucción	distancias	caracteres discretos
criterio de optimización	UPGMA Neighbour joining	X
criterio de agrupamiento	Evolución mínima	Máxima parsimonia Máxima verosimilitud

## Métodos de reconstrucción filogenética – una clasificación

### I.- Tipos de datos: distancias vs. caracteres discretos

- Los **métodos de distancia** primero convierten los alineamientos de secuencias en una matriz de distancias genéticas en base al modelo evolutivo seleccionado, la cual es usada por el método algorítmico de reconstrucción para calcular el árbol (**UPGMA y NJ**)
- Los **métodos discretos (MP, ML, Bayesianos)** consideran cada sitio del alineamiento (o una función probabilística para cada sitio) directamente



### Arboles filogenéticos: una introducción al bosque (I) terminología y conceptos básicos: anatomía de un árbol

- Definición: Un árbol filogenético es una estructura matemática usada para representar la historia evolutiva (relaciones de ancestro-descendiente) entre un grupo de secuencias o organismos. Dicho patrón de relaciones históricas es la estima hecha de la filogenia o árbol evolutivo.

**Anatomía básica de un árbol**

árbol no enraizado, sin direccionalidad

árbol enraizado, con direccionalidad, que indica relaciones ancestro-descendiente (((humano, chimp), gorila), orang)

- reconstrucción de caracteres ancestrales
- longitud de ramas
- soporte o confianza en splits

### Arboles filogenéticos: una introducción al bosque (II) enraizamiento de árboles

Tres métodos usados para el enraizado de árboles:

- grupo externo - (invertebrado) a grupo interno (vertebrados)
- b) punto medio – se pone la raíz en el punto intermedio del camino más largo del árbol
- c) duplicación génica – enraizamos en el nodo que separa a las copias parálogas

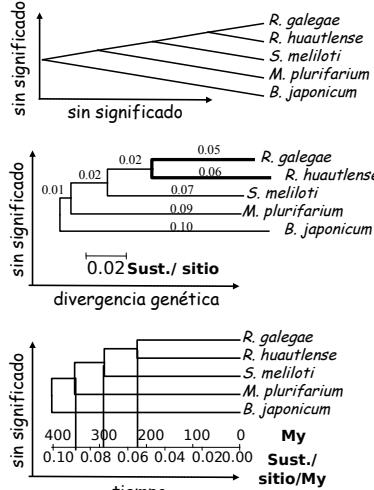
### Arboles filogenéticos: una introducción al bosque (III) terminología y conceptos básicos

- Los árboles son como móviles:** las ramas pueden rotarse sobre sí mismas sin afectar a las relaciones entre los OTUs; (((A,B),C),D),E) se puede representar como:

- Los árboles presentan distintos grados de resolución**

topología **estrella**      topología **parcialmente resuelta**      topología **totalmente resuelta**

### Arboles filogenéticos: una introducción al bosque (IV) terminología y conceptos básicos: tipos de árboles



- Un **cladograma**: sólo indica las relaciones de ancestría entre OTUs
- Una **topología aditiva** contiene la información sobre **longitudes de ramas**, que **refleja la distancia genética entre OTUs**. Así entre *R. galegae* y *R. huaultense* la distancia estimada es de:  $0.05 + 0.06 = 0.11$
- Una **topología ultramétrica**, dendrograma o árbol linearizado, representa un tipo especial de árbol aditivo en el que los nodos terminales son todas equidistantes de la raíz. Este tipo de árbol se emplea para representar el **tiempo evolutivo**, expresado bien como **años** o cantidad de divergencia medida por un **reloj molecular**

